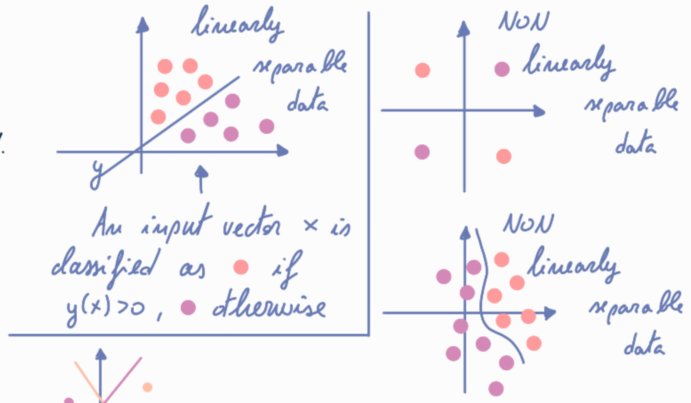
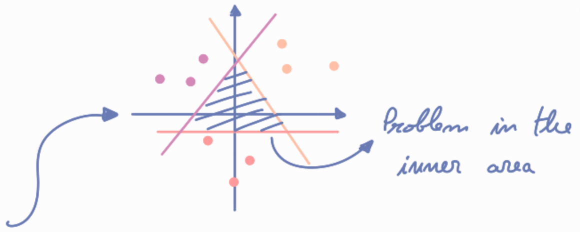
linearly separable data

Instances in a dataset are linearly separable iff there exists a hyperplane that separates the instance space in two regions, such that differently classified instances are separated. All the samples of one class should be on one side.
If the dataset is linearly separable there will exist infinite lines that can partition it.



The solution of the problem will be:

- $y(x) = w^T x + w_0$ for two classes
- $y_i(x) = w_i^T x + w_{0i}$ $\forall i \in [1, K]$ for K classes (we need to consider one linear model for each class)



N.B. this is similar to linear regression but we are not giving a probabilistic interpretation of the solution; y(x) is not anymore the prediction of the posterior probability of one class but it directly estimate the classification function

- $y(x) = w^T x + w_0 = \tilde{w}^T \tilde{x}$ with $\tilde{w} = \begin{pmatrix} w_0 \\ w \end{pmatrix}$, $\tilde{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$

$\tilde{W}$ contains all the parameters of the model

- $y(x) = \begin{pmatrix} y_1(x) \\ \vdots \\ y_K(x) \end{pmatrix} = \begin{pmatrix} \tilde{w}_1^T \\ \vdots \\ \tilde{w}_K^T \end{pmatrix} \tilde{x} = \tilde{W}^T \tilde{x}$ with $\tilde{W}^T = \begin{pmatrix} \tilde{w}_1^T \\ \vdots \\ \tilde{w}_K^T \end{pmatrix} = (\tilde{w}_1, \dots, \tilde{w}_K)$

directly the prediction of the output function (not prediction of the posterior)

This is called *linear model* because of course it is a linear combination. We are interested in computing $\tilde{W}$ (linear model in $\tilde{W}$).

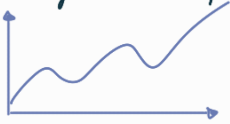
$$y(\tilde{x}, \tilde{W}) = \tilde{W}^T \phi(x)$$

We can apply a transformation on the input such that the model will no longer be linear in $\tilde{x}$ but it will still be linear in $\tilde{W}$. We require linearity with respect to $\tilde{W}$.

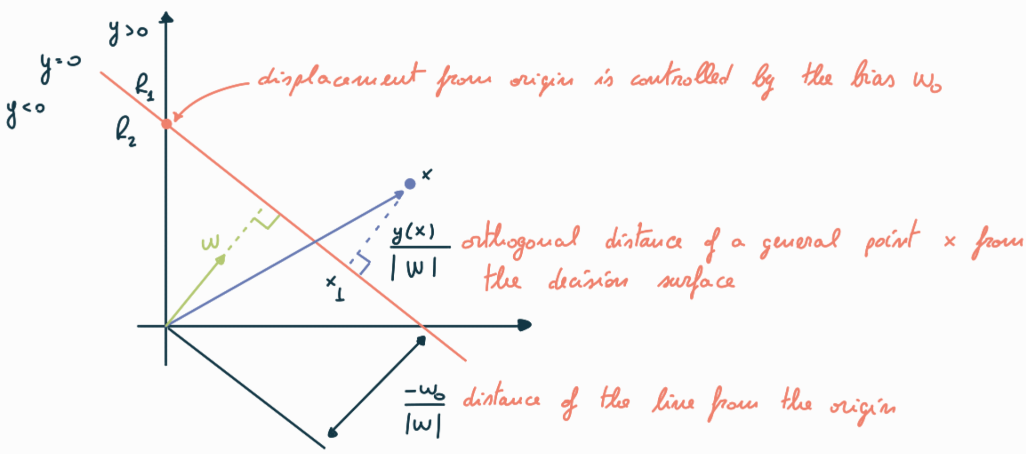
non linear x

e.g. $y(\tilde{x}, \tilde{W}) = w_0 + w_1 x + w_2 x^2 + w_3 x_3^3$

linear $\tilde{w}$

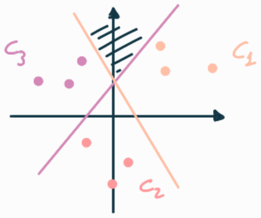


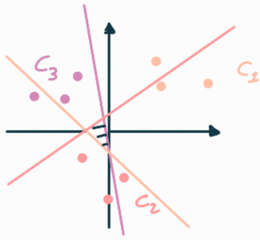
Geometric interpretation



Problem with multiple classes

we don't know how to predict the values in these regions





one VS the rest kind of classifier
C_i VS not C_i

- is c_3 or not ?
- is c_1 or not ?

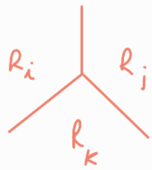
$K-1$ binary classifiers needed

one VS one kind of classifier
C_i VS C_j

- c_1 vs c_2 (does not care about c_3)
- c_2 vs c_3 (does not care about c_1)
- c_3 vs c_1 (does not care about c_2)

$K(K-1)/2$ binary classifiers needed

We can't reduce the problem of multi class classification to a set of binary classifiers. We need to consider a different model suitable for K classes, such that $\longrightarrow$ Prediction is possible in every region.



What we have

$D = \{(x_n, t_n)\}_{n=1}^{N}$ one-of-K coding scheme e.g. $t_n = (\overset{1}{0}, \dots, \overset{k-1}{0}, \overset{k}{1}, \overset{k+1}{0}, \dots, \overset{N}{0})^T \longleftrightarrow x_n \in C_k$

$t_j = 0 \ \forall j \in [1, \dots, N], j \neq k \ ; t_k = 1$

$y(x) = \tilde{W}^T \tilde{x}$ $\quad \tilde{x} = \begin{pmatrix} 1 \\ x \end{pmatrix}$ $\quad \tilde{W}^T = \begin{pmatrix} \tilde{w}_1 \\ \vdots \\ \tilde{w}_N \end{pmatrix}$ $\quad \tilde{w}_i = \begin{pmatrix} w_{0i} \\ w_i \end{pmatrix}$

Approaches to learn linear K-class discriminants

Approaches:
- Least squares
- Perceptron
- Linear Discriminant Analysis (LDA)
- Support Vector Machines (SVM)

Spoilers :

while methods such as logistic regression (probabilistic discriminative model) learn using the most representative samples for each class, SVMs learn using the most ambiguos and difficult to classify samples (the nearest to other classes) and use only them, ignoring the others.